Yupeng Tang

yupeng.tang@yale.edu | +1 4752018502| Github: charles-typ | Personal website: YP
51 Prospect Street, New Haven, CT

Research Interests

My research interests broadly include distributed systems, hardware accelerators, networking and their intersections.

Education

Yale University
Ph.D. in Computer Science — *Expected May 2026*

Xi'an Jiaotong University
Computer Science, Honors Science Program — *Expected May 2020*

University of California, Berkeley
Berkeley International Study Program — *August 2018 - May 2019*

Publications

- **Yupeng Tang**, Ping Zhou, et al. "*Exploring Performance and Cost Optimization with ASIC-Based CXL Memory*". **(Eurosys 24' Best paper Runner-Up)**.
- **Yupeng Tang**, Seung-seob Lee, Anurag Khandelwal. "*PULSE: Accelerating Distributed Pointer-Traversals on Disaggregated Memory*". **(Under Review)**.
- Hong Zhang, **Yupeng Tang**, Anurag Khandelwal, Ion Stoica. "*SHEPHERD: Serving DNNs in the Wild*". **(NSDI 23')**.
- Anurag Khandelwal, **Yupeng Tang**, Rachit Agarwal, Aditya Akella, Ion Stoica. "*Jiffy: Virtual Memory for Serverless Analytics*". **(Eurosys 22')**.
- Seung-seob Lee, Yanpeng Yu, **Yupeng Tang**, Anurag Khandelwal, Lin Zhong, Abhishek Bhattacharjee. "*MIND: In-Network Memory Management for Disaggregated Data Centers*". **(SOSP 21')**.
- Hong Zhang, **Yupeng Tang**, Anurag Khandelwal, Jingrong Chen, Ion Stoica. "*Caerus: NIMBLE Task Scheduling for Serverless Analytics*" **(NSDI 21')**.

Research Experience

Bytedance Inc — Dr. Zhou
Research Intern — *May 2024 - Present*

- **CXL 2.0/3.0 memory pooling**: I am rejoining the Bytedance Infrastructure Lab this summer to explore innovative CXL 2.0/3.0 memory pooling directions.

Bytedance Inc — Dr. Zhou
Research Intern — *June 2023 - Oct 2023*

- **Exploration of CXL memory**: Designed and implemented benchmarks to unveil performance characteristics of ASIC-based CXL memory modules. Identified use cases within Bytedance data centers and evaluated the performance of real data-center applications under various tiered memory management policies.

Yale University — Dr. Hong, Dr. Seung-seob, Prof. Anurag Khandelwal
Graduate Research Assistant — *2020 - Present*

- **Near-memory computing for disaggregated architectures**: Currently leading a project focused on in-network optimizations for irregular memory accesses in disaggregated data centers. The primary goal is to design a general offload interface for these accesses to remote memory, leveraging the programmability of SmartNICs.
- **MIND**: MIND is a rack-scale memory disaggregation system that uses programmable switches to embed memory management logic in the network fabric. My work involved analyzing and benchmarking cache coherency issues in traditional Distributed Shared Memory (DSM) systems.
- **Shepherd**: Shepherd is a DNN model serving system that achieves high system goodput and scalability while maximizing compute resource utilization. My work focused on designing and implementing the model serving system and scheduling algorithm.

- **Caerus**: Caerus is a task scheduling framework for Serverless platforms that efficiently pipelines task execution to minimize execution costs while achieving Pareto-optimal trade-offs between cost and job completion time (JCT) for arbitrary data analytics. My work focused on designing and implementing the task scheduling framework and exploring optimizations in data-dependent execution pipelines.

University of California, Berkeley Ph.D. Anurag Khandelwal, Prof. Ion Stoica
Undergraduate Research Assistant @ RISELAB *September 2018 - September 2020*

- **Jiffy**: Jiffy is a distributed memory management system that decouples memory capacity and lifetime from compute in the serverless paradigm. My work focused on designing and implementing reliable and elastic auto-scaling functionality for distributed memory management, along with several design optimizations to reduce data management latency.

Xilinx Labs Asia-Pacific Dr. Tuan Nguyen, Dr. Chengchen Hu
Research intern *Feb 2020 - May 2020*

- **Network Telemetry with FPGA-NICs**: Develop NIC drivers for custom FPGA-NICs and application frontend for presenting telemetry streaming data.

Microsoft Research Asia Dr. Lintao Zhang, Dr. Qi Chen
Research intern @ Intelligent Cloud and Edge Research Group *May 2019 - August 2019*

- **MLSystem**: Combine machine learning techniques with traditional operating systems structure to replace the abstraction layer and improve system performance.

TALKS

- In-Network Memory Management for Disaggregated Data Centers, in-person talk, **Bytedance Infrastructure Lab**, July 2023.
- Jiffy: elastic far-memory for stateful serverless analytics, virtual talk, **Huawei Technologies**, May 2022.
- NIMBLE Task Scheduling for Serverless Analytics, virtual talk, **Yale CSL seminar**, March 2021.

SKILLS

- **Languages**: C, C++, Python, Verilog, Rust, Matlab, TeX, Assembly(RISC-V), PHP, Go, Shell
- **Software/Hardware Development:** Git, Vim, Linux, CMake, Vivado, Vitis, DPDK, OpenCV, CUDA, Apache thrift, Pytorch
- **Hardware Accelerators**: FPGA, SmartNICs, Programmable Switch

TEACHING

- TA for CPSC 529 Principles of Computer System Design in Yale University 2022.09 - 2022.12
  - **Offload a Rust MoveNet inside Linux Kernel**: Design and implement a course project that enables a Rust kernel module to communicate with Linux Kernel camera module directly without crossing user-kernel boundary.
- TA for CPSC 538 Big Data Systems in Yale University 2021.02 - 2021.05
- TA for Progamming Fundamentals in Xi'an Jiaotong University 2019.09 - 2019.12

SELECTED HONORS & AWARDS

- Yale Student Fellowship 2020.09
- Third Prize in Group Programming Ladder Tournament (National) 2018.03
- PengKang Scholarship, Xi'an Jiaotong University (Intra-school, **top 5%**) 2017.09
- Scholarship under the State Scholarship Fund (National) 2018.05